Data 583 Life Expectancy (WHO)

Justin Chan, Kenny Tong, Viji Rajagopalan

22 Mar, 2023

Data 583 Life Expectancy - Final Report (Life Expectancy Data)

1. Introduction and Hypotheses

Life expectancy has always been an area of interest for humanity. The goal of this project is to study a dataset that contains information on life expectancy and identify some of the variables that significantly impact life expectancy. The dataset chosen for the study has life expectancy data of 193 countries between 2000-2015, together with different predictive factors. Broadly speaking, predicting variables are categorized into 4 major areas : Immunization, Mortality, Economical, and Social, containing a total of 21 individual variables. Our hypothesis is that a subset of variables from this dataset would be able to explain and predict life expectancy with good accuracy (say > 80%). The dataset has a mix of variable types – continuous and discrete. Within discrete types, some variables are ordinal, and some are non-ordinal or nominal. With such a mix and complexity of data, we also hypothesize that all variables will not share a simple linear relationship with the predictor variable and modelling of life expectancy will require a more complex model. We analyze and validate several statistical models throughout the report with the primary goal of identifying an adequate model for predicting life expectancy.

2. Dataset overview

Variables Summary and Categories

Life expectancy is the response variable in this dataset. This represents the mean life expectancy (in age) by specific country and year combination. Refer Figure-1 below for the list of predictor variables and their categories.

The dataset also had 2563 missing values in various columns and ..... the NA values are generally imputed by the respective column mean. Percentage expenditure variable is removed from the entire assessment as the values present in this column are unclear. Another variable country is removed for the purpose of studying life expectancy globally.

#(Kenny, please add if any other deletion done in .... line and fill in on x and y)

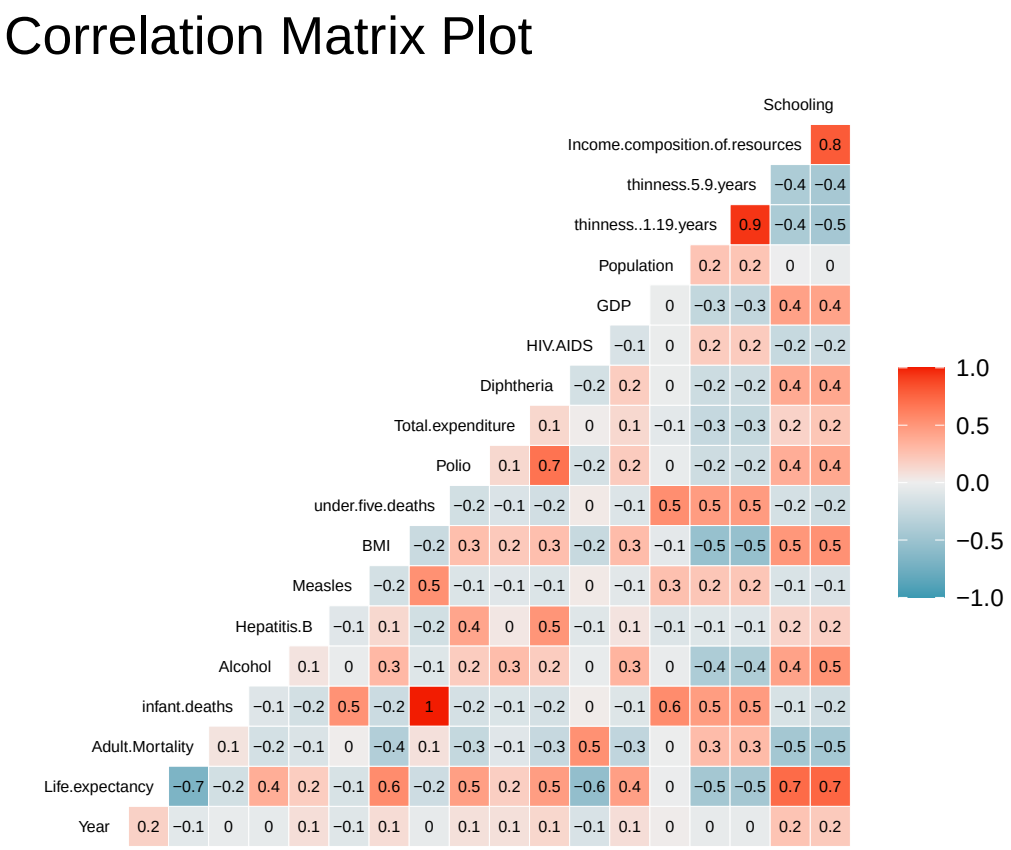
Variable	Unit of Measurement/Data Category	Continuous vs Discrete	Variable	Unit of Measurement/Data Category	Continuous vs Discrete
Life Expectancy	Years Old (Age)	Continuous	Total expenditure	Percentage	Continuous
Country	Nominal Data	Discrete	Percentage expenditure	Percentage	Continuous
Year	Ordinal Data	Discrete	GDP	Currency (USD)	Continuous
Status	Nominal Data	Discrete	Population	Count	Discrete
Adult Mortality	Count Data	Discrete	Income composition of resources	Percentage	Continuous

Variable	Unit of Measurement/Data Category	Continuous vs Discrete	Variable	Unit of Measurement/Data Category	Continuous vs Discrete
Infant deaths	Count Data	Discrete	Schooling	Mean (Years)	Continuous
Under-five deaths	Count Data	Discrete	Alcohol	Litres	Continuous
Hepatitis B	Percentage	Continuous	HIV/AIDS	Percentage	Continuous
Measles	Count Data	Discrete	BMI	Average BMI	Continuous
Polio	Percentage	Continuous	Thinness 1-19 years	Percentage	Continuous
Diphtheria	Percentage	Continuous	Thinness 5-9 years	Percentage	Continuous

Following table shows how the above variables are grouped based on the 4 categories.

Data Categories	**Variables**
Economical Data	Total expenditure, Percentage expenditure, GDP, Income composition of resources
Social Data	Country, Status, Population, Schooling, Alcohol, BMI, Thinness 1-19 years, Thinness 5-9 years
Mortality Data	Adult Mortality, Infant deaths, Under-five deaths
Immunization Data	Hepatitis B, Measles, Polio, HIV/AIDS, BMI, Diphtheria

The resulting dataset with x rows and y columns are then studied closely to understand their correlation effects with the response variable life expectancy. Following is a correlation matrix on all the variables in the dataset.



It can be noted that the response variable life expectancy is highly correlated with income composition, schooling and adult mortality variables with a correlation value of 0.7, 0.7 and −0.7. Life expectancy is moderately correlated

with variables BMI, Polio, Diphtheria, HIV.AIDS and thinness variables with a correlation value of 0.6, 0.5, 0.5, −0.6, −0.5. Broadly, it can be noted that the two categories of the variables appear to be highly correlated with life expectancy when compared to other categories.

Initial analysis using linear regression

Life expectancy is a continuous variable and the first choice is building a linear regression model which is simple and interpretable. A BIC backward step model variable selection method is also applied on the full model to arrive at a parsimonious model containing only significant predictor variables. Following table Table A provides a summary of the two models.

Models	No. of Variables	AIC Score	Adj R-squared Score
Original Model	20	7642.14	0.8299
Reduced Model	12	7604.24	0.8296

The number of independent variables are now effectively reduced to 12, together with a lower AIC score of 7604.34. Meanwhile, the adjusted R-squared score is well kept at nearly the same level as in the original model. The reduced model is able to explain more than 82% of variation in the response variable and its performance is above the anticipated 80%. The reduced model now contains the following 12 variables : Adult.Mortality + infant.deaths + Hepatitis.B + BMI + under.five.deaths + Polio + Diphtheria + HIV.AIDS + GDP + thinness..1.19.years + Income.composition.of.resources + Schooling.

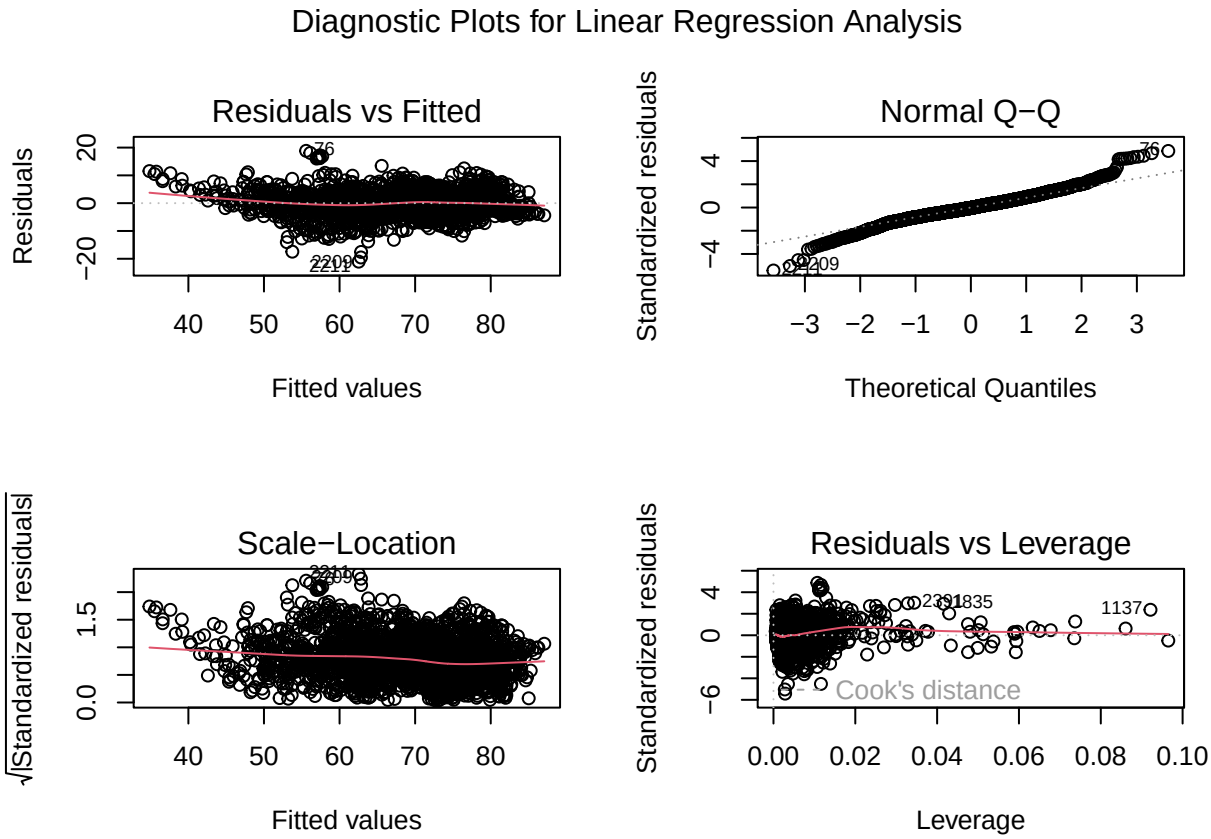
The reduced model from this step is selected as the first model for the dataset. It is used for further evaluation from a linear model stand point and will be referred to as linear model in the remainder of the discussion.

3. Regression Analysis

Linear model and diagnostics

The linear model shows that we are able to explain approximately 82% of variability of our response variable using the selected predictor variables. The next step is to look at the error diagnostics from the model.

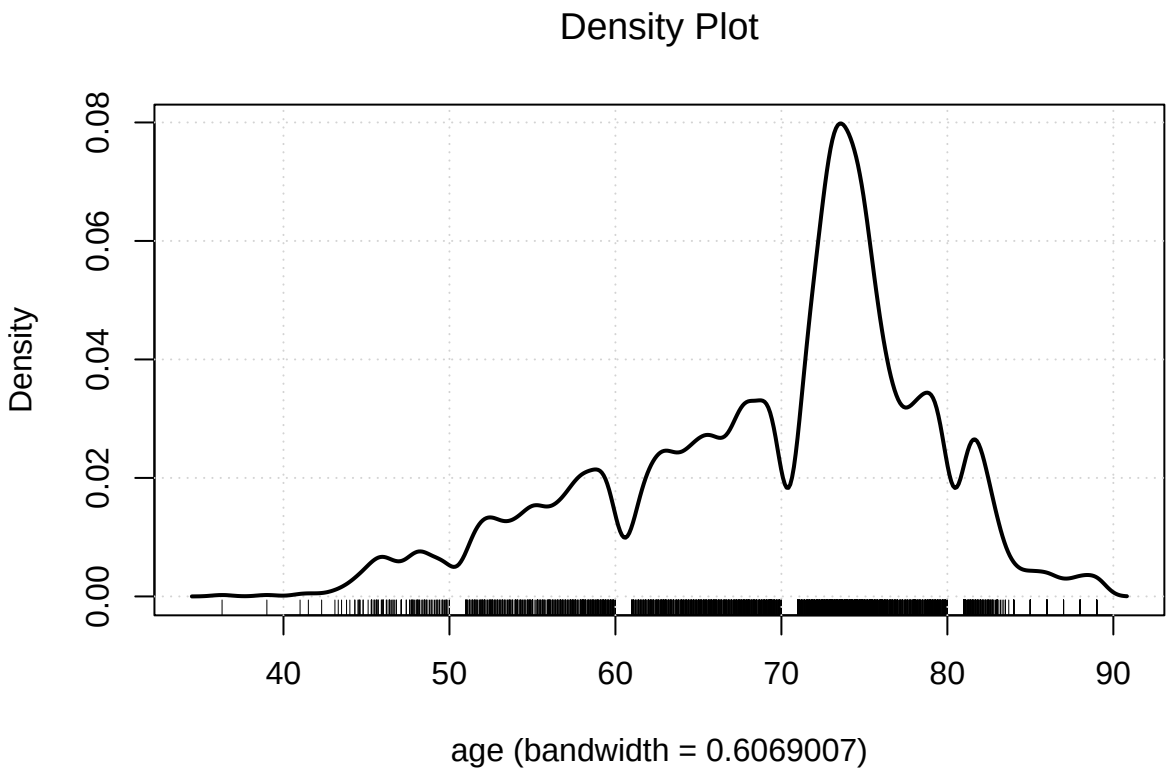
```
par(mfrow=c(2,2))

plot(lmmod2)
mtext("Diagnostic Plots for Linear Regression Analysis", side = 3, line = -1, outer = TRUE)
```



The QQ plot suggests that the model is heavy tailed and the data on both ends of the quantiles do not fit on a straight line. This is an indication that the current linear regression based model is not fitting the data well. Based on this, we undertake additional testing to validate if the model is adequate and valid.

```
densityPlot(~ Life.expectancy, show.bw=TRUE, method="kernel", data = df, xlab="age")
title(main="Density Plot",font.main= 1)
```

a. Life expectancy variable distribution



From the Density plot above, it can be seen that the mean of the distribution isn't symmetrical and the mean isn't centered at 0, indicating the response variable life expectancy is not normally distributed. From the Normal QQ plot of the regression model, it can be observed that there is a distinct curve in the middle of plot rather than a straight line, this indicates that there could be a bimodal distribution to response variable. These observations necessitate validation of distribution of the response variable. One of the assumptions of linear model is that the response variable is normally distributed and if it is proven that the variable is not normally distributed, linear regression model for this dataset will be invalid due to violation of assumptions.

b. Normal distribution test for response variable Shapiro-Wilk test is a statistical test for normality and a p-value that is very small and is less than 0.05 proves the variable in consideration is not normally distributed. Shapiro-Wilk test on the response variable life expectancy resulted in a p-value of $<$ 2.2e-16 as seen in the test output below.

```
##
##  Shapiro-Wilk normality test
##
## data:  df$Life.expectancy
## W = 0.95676, p-value < 2.2e-16
```

This proves the response variable is not normally distributed and additionally, hypothesis tests for validating correct specification of parametric MLR models are conducted to identify if the selected linear model specification is valid for the given dataset.

c. Parametric model specification test Ramsey's RESET test is a test conducted to validate the correctness of the functional form. A p-value that is very small and is less than 0.05 rejects that the functional form is correctly specified. RESET test is conducted on the linear model and the resulting p-value as seen from the output below is $< 2.2e-16$.

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.2.2
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.2.2
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
resettest(lmmod2)
```

```
##
##  RESET test
##
## data:  lmmod2
## RESET = 121.2, df1 = 2, df2 = 2763, p-value < 2.2e-16
```

Based on the test, the linear model is rejected as the correct functional form for modelling the underlying data.

d. Consistent nonparametric inference The consistent nonparametric inference test is a hypothesis test for correct specification of parametric MLR models. This allows to estimate if the functional for given parameter estimates is reasonable when compared. A p-value that is very small and is less than 0.05 rejects that the functional form for given parameter estimates is reasonable.

```
##
## Consistent Model Specification Test
## Parametric null model: lm(formula = Life.expectancy ~ Adult.Mortality +
##                           infant.deaths + Hepatitis.B + BMI + under.five.deaths
##                           + Polio + Diphtheria + HIV.AIDS + GDP +
##                           thinness..1.19.years +
##                           Income.composition.of.resources + Schooling, data =
##                           df, x = TRUE, y = TRUE)
## Number of regressors: 12
## IID Bootstrap (399 replications)
##
## Test Statistic 'Jn': 21.17521    P Value: < 2.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Null of correct specification is rejected at the 0.1% level
```

As noted from above table, the p-value for linear model is $< 2.22e-16$ and this output suggests that the linear model is rejected.

All the diagnostic tests indicate that linear regression is not an appropriate model for the given data. This proves one of the hypothesis of our project that a simple linear model may not be adequate in explaining the variability in the response variable life expectancy.

Parametric regression models and relative assessments

As the linear model is not adequate, we move on to other parametric regression models that do not assume normal distribution and known to perform well on complex and mixed data types. There are several models that could be used in the assessments and we selected LASSO and Neural Net with linear activation function for the given dataset.

The following variables are selected for rest of the modeling based on correlation of the variables with the response variable and our knowledge on the domain. Here is a summary of the variable selection and comments describing the reason for removal of the variable.

Column Name	Type	LM	LASSO	NN	NPREG	Reason of Removal
Country	(Discrete)					Since we wanted to build models for all countries
Year	(Discrete)					ordinal type data and based on domain knowledge, not consider important
Status	(Discrete)					nominal type data and based on domain knowledge, not consider important
Adult Mortality	(Discrete)	X	X	X	X	
Infant deaths	(Discrete)	X	X	X	X	
Under-five deaths	(Discrete)	X	X	X	X	
Hepatitis B	(Continuous)	X	X	X	X	
Measles	(Discrete)					Since it is a count and discrete type data and weak correlation with our predictor
Polio	(Continuous)	X	X	X	X	
Diphtheria	(Continuous)	X	X	X	X	
Total Expenditure	(Continuous)					based on domain knowledge, not consider important
Percentage Expenditure	(Continuous)					based on domain knowledge, not consider important
GDP	(Continuous)	X	X	X	X	
Population	(Discrete)					no correlation with our predictor indicated by our correlation plot
Income composition of resources	(Continuous)	X	X	X	X	
Schooling	(Continuous)	X	X	X	X	
Alchol	(Continuous)					based on domain knowledge, not consider important
HIV/AIDS	(Continuous)	X	X	X	X	
BMI	(Continuous)	X	X	X	X	
Thinness 1-19 years	(Continuous)	X	X	X	X	
Thinness 5-9 years	(Continuous)					range already covered in 1-19 Thinness 1-19 years
status.val	(Continuous)					based on domain knowledge, not consider important

First, the dataset is divided into a train and test datasets with an approximate 70%-30% of the complete data using a random sampling process so long run performance of the models can be estimated. The train dataset has 2000 records and test dataset has 778 records in total.

Three types of models using linear regression (LM), LASSO and NeuralNet with linear activation function are built on the *train* dataset and the PRESS (Predicted Residual Error Sum of Squares) statistic is calculated using *test*

to identify the best performing model. Following table summarizes and compares the PRESS statistics.

	LM	LASSO	NN
PRESS	15.93367	15.98972	22.43056

Based on the output, it can be seen that LM and LASSO models perform better than NeuralNet model for this dataset. P.S: Linear model is used in these assessments for benchmarking purposes and not for actual use as the model is not valid.

Selecting LM and LASSO, the R^2 is also measured for the models. It can be seen that LASSO performs nearly at the same level as the linear model (LM).

	LM	LASSO
R2	0.829139273435324	0.829061931452822

Based on these assessments, LASSO is a viable model that can be considered for this dataset that has approximately 82 for predicted R^2 value and meets the performance goals expectations.

Diagnostics

Nonparametric regression

Nonparametric regression is considered as another good option for the complex and mixed dataset that is of interest here due to proven flexibility and adaptability nature of these models. One important difference between nonparametric model and rest of the parametric models is that the entire data is used in the model training process. The nonparametric regression is carried out with local linear estimator and cv.aic for automated bandwidth selection. This bandwidth selection method specifies expected Kullback-Leibler cross-validation (Hurvich, Simonoff, and Tsai (1998))and in general provides consistent estimates.

The output of the nonparametric regression model indicates an R^2 value of 87% approximately. This is the summary measure of in-sample fit for the model lies in the range of [0,1]. 1 denotes a perfect fit to the sample data and 0 indicates no fit. This is the counterpart to R^2 of linear model.

We acknowledget that the R^2 for LASSO model is calculated based on a train vs test setup and nonparametric regression R^2 is calculated on the complete dataset and a nonparamertic train-test fitting is identified as a future scope item that will be looked into. As the nonparametric model is a cross validated model and can provide long run performance, the LASSO and nonparametric model coeffients are compared and summarized in the table below.

	NPREG	LASSO
R2	0.8722143	0.829061931452822

The nonparametric model has a higher R^2 of 87%(approx) when compared to the parametric model R^2 of 83%(approx). So, it is concluded that nonparametric model fits the given dataset better and is selected among the assessed models for use.

In order to arrive at a more parsimonious model, significance of the variables used in nonparametric regression is measured and per below output all the variables are significant and will be retained in the model.

```
#npsigtest_npreg <- npsigtest(model_np)    #10 HRs to run...
```

Diagnostics

```
> npsigtest(model_np)

Kernel Regression Significance Test
Type I Test with IID Bootstrap (399 replications, Pivot = TRUE, joint = FALSE)
Explanatory variables tested for significance:
Adult.Mortality (1), infant.deaths (2), Hepatitis.B (3), BMI (4), under.five.deaths (5), Polio (6), Diphtheria (7), HIV.AIDS (8),
 GDP (9), thinness..1.19.years (10), Income.composition.of.resources (11), Schooling (12)

             Adult.Mortality infant.deaths
Bandwidth(s):       389457535       6733757
             Hepatitis.B      BMI under.five.deaths
Bandwidth(s):   225092161 79216285          95308072
               Polio Diphtheria HIV.AIDS
Bandwidth(s): 5825954   19248839 1.393258
                    GDP thinness..1.19.years
Bandwidth(s): 167351562078             37667667
             Income.composition.of.resources
Bandwidth(s):                         1071202
             Schooling
Bandwidth(s):  15165344

Individual Significance Tests
P Value:
Adult.Mortality                 < 2e-16 ***
infant.deaths                   < 2e-16 ***
Hepatitis.B                     0.047619 *
BMI                             < 2e-16 ***
under.five.deaths               < 2e-16 ***
Polio                           < 2e-16 ***
Diphtheria                      < 2e-16 ***
HIV.AIDS                        < 2e-16 ***
GDP                             < 2e-16 ***
thinness..1.19.years            < 2e-16 ***
Income.composition.of.resources < 2e-16 ***
Schooling                       < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 1: npsigtest_npreg result

4. Conclusion

The non-normality nature of the dataset has been observed and verified by rigorous testings and validation in this report. This characteristic has greatly limited the applicablity of many popular common models which rely on the assumption of normal distribution. After further model assessment, we are finally able to come into the best available conclusion that NOREG and LASSO are the two best-performing models based on model performance indicators like MSE and R-squared values.

We find that life expectancy is....[KT : I literally have no idea what to put there....may be I leave it to Viji to continue and finally conclude the report :)]

Future improvements

While a number of different models and statistical tests have been explored within a limited time frame of this project, we can hardly conclude we have identified the globally optimal models. In fact, in order to limit the complication of this analysis and make it reasonably achievable, we have adopted certain model and analysis simplification in a few aspects. These assumptions/simplification may, however, potentially have adverse effect on our underlying models accuracy. As rooms of further improvement works based on this report, the following aspects are suggested for future exploration, studies and implementation to see if an even better-performing model can be attained.

1. Currently, no particular handling has been done to process the categorical, ordinal, and nominal variables. Current variables are simply fit into different models with "as-is" data basis. Further exploration on whether some techniques (such as Variables Encoding/transformation, factorization factor()) can be deployed to achieve models improvement is preferable.

2. Performing non-parametric model in our analysis has taken a substantial amount of computing resources. The studies on the non-parametric model what we have achieved so far is generally sufficient for measuring long run performance. While resources and time allow in the future, we may consider performing further

fine-tuning on this by enforcing dataset splitting into training and testing set under non-parametric model fitting, which can possibly have a better account of the model performance.

3. According to the earlier Multicollinearity studies (Part 2), correlation is found between the variables infant.deaths and under.five.deaths. It is understood that such correlation may cause undesirable effect on model accuracy, fitting and interpretation. To resolve this issue, we may explore possible tactics such as removing one of the correlated variables, or using factor analysis (factanal) to address the multicollinearity issue to enhance the models.

4. Currently in our analysis, data implantation (rather than removing the records with NA values) has been deployed in order to retain as many records as possible and simplify/streamline the subsequent analysis. Although data implantation is a common industry practice, We are not 100% sure if such procedure would affect the model accuracy. In this regard, we may investigate and compare different null data handling techniques (apart from data implantation using mean) and investigate if we can achieve our modelling improvement as a result.

Interesting challenges

During this project, a number of challenges are encountered. These challenges have created extra hurdles and unforeseeable overheads on our projects, or have caused unexpected complication for the project team in order to efficiently and confidently identify the most suitable models.

1. Running npreg on our model is extremely time-consuming. It took 30 hours in a notebook computer. This undesirable situation has seriously constrained our flexibility in fine-tuning and re-running the model with different model settings such as variable combination because we simply cannot afford adjusting the model fitting to look more a potentially more optimal model fitting.

2. Similarly, running model significance took more than 30 hours. This has caused similar consequence as the previous point 1.

3. As mentioned in the earlier analysis, bimodal distribution is identified in the dataset, which has violated the basic assumptions of many parametric models. This behavior has therefore severely limited the applicability of many parametric modelling. We also lack of sufficient knowledge on how to optimally model and analyze bimodal distribution.

4. The dataset has demonstrated quite a high proportion of NA values. Several columns contain significantly more than 5% of NA values. If we decide to adopt the 5% threshold and remove all records (e.g. dropping columns or removing rows) with NA values which exceeds the 5% threshold, it would result in a significant amount of records being removed and only remain a much smaller sample size available for further analysis. This may tremendously and adversely impact and deteriorate the analysis accuracy and reliability.

Appendix

Checking for Multicollinearity

As Multicollinearity can potentially affect the accuracy of regression model and we have 22 variables, a correlation study is undertaken to understand and assess the situation. A correlation plot has identified a number of correlation problems. It is found that infant deaths and under.five.deaths are nearly 100% correlated. The relation between the deaths rates of the two close age groups is easily interpretable. In addition, there are three heavily correlated pairs which is defined by the abs(correlation coefficient)>0.7 between the variables. They include (a) (immunization rate of) 'Polio'-vs-'Diphtheria', (b) 'income composition of resources'-vs-'Schooling', and (c) between the two thinness measures for the age groups 5-9 vs 10-19. Pairs (a) and (c) are justifiable while the relation for (b) demonstrate a relatively subtle relation. Other than that, the degree of multicollinearity is acceptable and not too worrying.